The Proximal Origin of SARS-CoV-2

Novel 2019 coronavirus

nCoV-2019 Evolutionary History

arambaut #1 February 17, 2020, 4:02pm

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

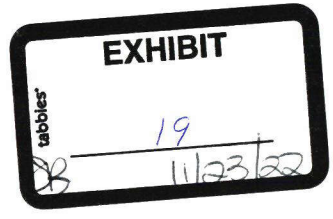
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

[*]**Corresponding author:**

Kristian G. Andersen
Department of Immunology and Microbiology,
The Scripps Research Institute,
La Jolla, CA 92037,
USA.



Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modelling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (Figure 1a). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (Figure 1a). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction

is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (Figure 1b)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (Figure 1b). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11-13]. The acquisition of polybasic cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a “mucin-like domain” that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein. Biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

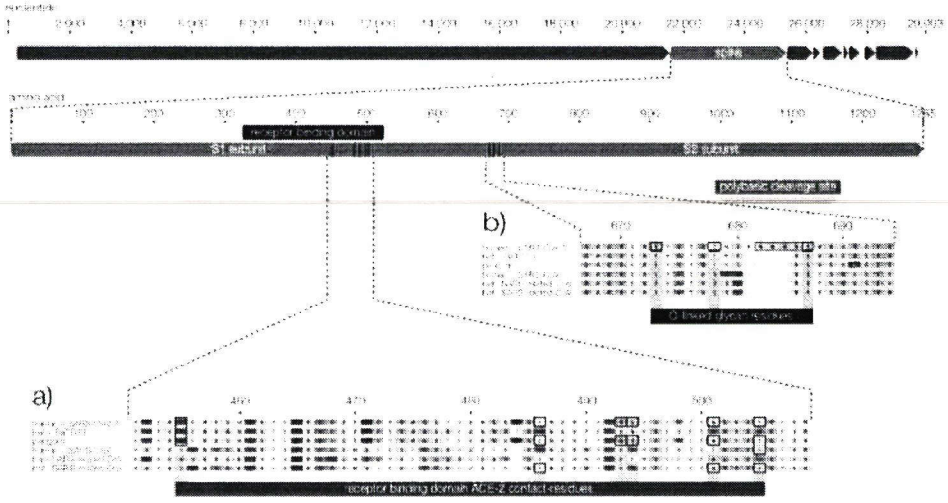


Figure 1. (a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. (**b) Acquisition of polybasic cleavage site and O-linked glycans.** The polybasic cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the polybasic cleavage site and O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146, MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[18,19].

Theories of SARS-CoV-2 origins

It is unlikely that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[17]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host. As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry

viruses that are genetically very similar to SARS-CoV-1 or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[18,19]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (Figure 1). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans. It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site, and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[20,21], compatible with the earliest retrospectively confirmed cases[22]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[23], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[24]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior

infection with SARS-CoV-1 or -2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

Selection during passage. Basic research involving passage of bat SARS-like coronaviruses in cell culture and/or animal models have been ongoing in BSL-2 for many years in multiple laboratories across the world[25-28]. There are also documented instances of the laboratory acquisition of SARS-CoV-1 by laboratory personnel working under BSL-2 containment[29,30]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[31]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro* .

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data will help resolve this issue. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (**https://www.gisaid.org/**) and contributed analyses and ideas to **Virological.org** (**http://virological.org/**). We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is supported by NIH grant 1U19AI135995-01. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS).

1. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.

2. Wu, F. et al., A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. Letko, M. & Munster, V. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.

4. Hoffmann, M. et al., The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv* 2020.01.31.929042 (2020) doi:10.1101/2020.01.31.929042.

5. Sheahan, T. et al., Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. *J. Virol.* **82** , 2274–2285 (2008).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17** , 181–192 (2019).

7. Wrapp, D. et al. Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. bioRxiv 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan coronavirus: deja vu. ***Virological.org* Analysis of Wuhan Coronavirus: Deja Vu** (2020).

9. Coutard, B. et al., The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350** , 358–369 (2006).

11. Longping, V. T., Hamilton, A. M., Friling, T. & Whittaker, G. R. A novel activation mechanism of avian influenza virus H9N2 by furin. *J. Virol.* **88** , 1673–1683 (2014).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28** , 19–38 (2009).

13. Luczo, J. M. et al., Evolution of high pathogenicity of H5 avian influenza virus: haemagglutinin cleavage site selection of reverse-genetics mutants during passage in chickens. *Sci. Rep.* **8** , 11518 (2018).

14. Ito, T. et al., Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75** , 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64** , 3297–3303 (1990).

16. Shengqing, Y. et al., Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301** , 206–211 (2002).

17. Menachery, V. D. et al., A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21** , 1508–1513 (2015).

18. Liu, P., Chen, W. & Chen, J.-P. Viral metagenomics revealed sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*) *Viruses* **11** , 979 (2019).

19. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

20. Phylodynamic analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. ***Virological.org*** **Phylodynamic Analysis | 93 genomes | 15 Feb 2020** (2020).

21. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. ***Virological.org*** **Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time** (2020).

22. Huang, C. et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-530183-5).

23. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders-Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52** , 986–987 (2003).

24. Wang, N. et al., Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33** , 104–107 (2018).

25. Ge, X.-Y. et al., Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503** , 535–538 (2013).

26. Hu, B. et al., Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13** , e1006698 (2017).

27. Zeng, L.-P. et al., Bat Severe Acute Respiratory Syndrome-like coronavirus WIV1 encodes an extra accessory protein, ORFX, involved in modulation of the host immune response. *J. Virol.* **90** , 6573–6582 (2016).

28. Yang, X.-L. *et al.,* Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of Severe Acute Respiratory Syndrome coronavirus. *J. Virol.* **90** , 3253–3256 (2015).

29. Lim, P. L. et al., Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350** , 1740–1745 (2004).

30. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3** , 679 (2003).

31. Letko, M. et al., Adaptive evolution of MERS-CoV to species variation in DPP4. *Cell Rep.* **24** , 1730–1737 (2018).

profbillg1901 #2 February 17, 2020, 3:51pm

I concur with this analysis at the amino acid level. I would add, however, that even if two virus isolates have an IDENTICAL amino acid sequence, that would not automatically make one the origin of the other. In my post 10 days ago, **Tackling Rumors of a Suspicious Origin of nCoV2019** , I showed that mutations in the 3rd wobble base in an otherwise conserved region accumulate as to make seemingly identical viruses very much separated in time – indeed over decades.

It is time to leave playing “who’s ur daddy” to daytime television. No known viral RNA sequence is the daddy of SARS-CoV-2. No match. If it doesn’t fit, you must acquit. It is all bat guano. Whatever analogy you like. A thrust of distrust and recrimination is not the path to protecting the planet from a viral catastrophe.

Bill Gallaher